Self-Influence Governs Generalization: A von Mises Expansion Approach

author names withheld

Under Review for the Workshop on High-dimensional Learning Dynamics, 2026

Abstract

We study generalization through the lens of self-influence: how strongly a training example affects the learned predictor and its associated losses. Using a second-order von Mises expansion of the loss functional, we derive leading-order influence-based estimators for the expected generalization gap. For i.i.d. samples, the expected gap is governed by average self-influence. For hierarchical sampling, additional same-parent cross-influence terms appear, providing a mechanism by which augmentations and correlated views affect generalization. Empirically, the resulting estimators accurately track generalization in well-generalizing regimes and become conservative near memorization, providing a diagnostic signal for the breakdown of the leading-order approximation.

Keywords: Generalization, Influence Functions, Hierarchical Sampling, von Mises Expansion

1. Introduction

Generalization reflects how learned predictors respond to finite-sample fluctuations. We study this response through a von Mises expansion of the loss functional, treating the empirical distribution as a perturbation of the population distribution. In this work, we develop the von Mises expansion [5] of the loss functional, treating the empirical data distribution as a perturbation of the true distribution. Such an expansion, to first order, was popularized by Hampel [6] in the robustness literature for measuring the influence of an individual data point on an estimator. Koh and Liang [9] developed this into an important tool in modern-day machine learning, by estimating the influence of one data point on the loss of another. To get to expectations of the empirical generalization gap, we need to push the von Mises expansion to second order. The expressions involve kernels in the data space that we designate as influence kernels.

Our contributions are:

- A second-order influence-function expansion for the expected generalization gap.
- A hierarchical extension showing how same-parent cross-influence modifies generalization.
- Empirical demonstrations in high-dimensional linear regression, including conservative test-based estimators near memorization.

2. Related Work.

Our work builds on von Mises expansions and asymptotic M-estimation theory [1, 5, 10, 14–16], while adopting the influence-function perspective popularized in machine learning by Koh and

Liang [9]. Unlike prior work emphasizing interpretability or data attribution, we focus on influence geometry as a probe of generalization, including hierarchical sampling settings motivated by modern self-supervised learning [2, 13].

3. Preliminaries

For a data distribution Π, define the corresponding loss (see examples in Appendix A):

$$L(\theta, \Pi) = \int l(z, \theta) d\Pi(z).$$

We could extend the definition to signed measures Π under appropriate conditions.

Let $(Z_1, Z_2, \ldots, Z_N)$ be N exchangeable [4] random samples with the marginal distribution of Z_i being P. This sample gives us an empirical distribution: $\hat{P} := \frac{1}{N}\sum_i \delta_{Z_i}$, where δ_z is the Dirac measure centered at z. We will treat $\hat{P}$ as a random measure.

We consider ridge-regularized empirical risk minimization:

$$\hat{\theta}_\lambda = \arg\min_\theta \left(L(\theta, \hat{P}) + \frac{\lambda}{2}\|\theta - \theta_{\text{ref}}\|^2 \right),$$

with corresponding population optimum θ^*_λ.

The test loss influence function of Koh and Liang [9] motivates us to define an influence kernel. Let $H^* := \nabla^2_\theta L(\theta^*_\lambda, P)$.

Definition 1 *(Influence Kernel) The influence kernel $\mathcal{K}_\lambda(z, z')$ is defined as*

$$\mathcal{K}_\lambda(z, z') := \nabla_\theta l(z, \theta^*_\lambda)^\top (H^* + \lambda \mathbb{I})^{-1} \nabla_\theta l(z', \theta^*_\lambda).$$

Relative to the definition of the loss influence function in Koh and Liang [9] (see also Eq. (4) in Appendix C.1), we drop the negative sign and evaluate the loss under the true distribution P at the ideal optimum parameter $\theta = \theta^*_\lambda$.

4. Approach

4.1. von Mises Expansion Approach

We interpolate between the population and empirical distributions via

$$\hat{P}(\epsilon) = (1 - \epsilon)P + \epsilon\hat{P},$$

and denote the corresponding minimizer by $\hat{\theta}(\epsilon)$.

4.2. Optimization for $\hat{P}(\epsilon)$

For the sake of notational simplicity, from this point on, we suppress the λ-dependence of optimal parameters θ^*, $\hat{\theta}$, and $\hat{\theta}(\epsilon)$. Optimality requires that $\nabla_\theta L(\hat{\theta}(\epsilon), \hat{P}(\epsilon)) + \lambda(\hat{\theta}(\epsilon) - \theta_{\text{ref}}) = 0$. As ϵ varies from zero to one, $\hat{\theta}(\epsilon)$ interpolates between θ^* and $\hat{\theta}$, as in Fig. 4 in the Appendix B.

4.3. ϵ-Generalization Gap

Notice that $L(\hat{\theta}(\epsilon), P) - L(\hat{\theta}(\epsilon), \hat{P}(\epsilon))$, at $\epsilon = 1$, is the generalization gap $\Delta_{\text{gen}}(\hat{\theta}) := L(\hat{\theta}, P) - L(\hat{\theta}, \hat{P})$. Let us call this quantity for any ϵ the ϵ-generalization gap:

Definition 2 (ϵ-generalization gap) *We define ϵ-generalization gap as*

$$\Delta_{\epsilon-\text{Gen}}(\theta) := L(\theta, P) - L(\theta, \hat{P}(\epsilon)).$$

Note that, at $\epsilon = 1$, this definition matches the conventional definition of the generalization gap [7] for functions parametrized by $\theta \in \Theta$. We now shift focus from $\Delta_{\text{gen}}(\hat{\theta})$ to $\Delta_{\epsilon-\text{Gen}}(\hat{\theta}(\epsilon))$, especially for small ϵ, which makes it a more tractable endeavor. For that purpose, we Taylor expand the ϵ-generalization gap at the (ϵ and $\hat{P}$-dependent) optimal point $\hat{\theta}(\epsilon)$ around $\epsilon = 0$.

Proposition 3 *Under Assumptions 10 and 11,*

$$\Delta_{\epsilon-\text{Gen}}(\hat{\theta}(\epsilon)) = -\epsilon L(\theta^*, \hat{P} - P) + \epsilon^2 \int \mathcal{K}_\lambda(z, z') d(\hat{P} - P)(z) d(\hat{P} - P)(z') + o_P(\epsilon^2). \quad (1)$$

For proof, see Appendix F.1. For i.i.d. samples, the first term fluctuates as $O_P(N^{-1/2})$, while the second contributes at $O_P(N^{-1})$ to expectations. When we compute the expected value of the ϵ-generalization gap, the first term contributes zero.

4.4. Main Theorems

Our main results (see Appendix G for proof):

Theorem 4 *In the i.i.d. samples case, we have*

$$\mathbb{E}_{\hat{P}}[\Delta_{\epsilon-\text{Gen}}(\hat{\theta}(\epsilon))] = \frac{\epsilon^2}{N}\Big[\mathbb{E}_{Z\sim P}\big[\mathcal{K}_\lambda(Z, Z)\big] - \mathbb{E}_{(Z,Z')\sim P^{\otimes 2}}\big[\mathcal{K}_\lambda(Z, Z')\big]\Big] + o(\epsilon^2). \quad (2)$$

For $\lambda = 0$, the off-diagonal term vanishes.

However, suppose we have the following hierarchical sampling setting. Sample $(X_1, X_2, \ldots, X_n)$ independently from P, and then sample $(\Xi_1, \Xi_2, \ldots, \Xi_N)$ independently from $(\frac{1}{n}\sum_{\alpha=1}^{n} \delta_{X_\alpha})$. Then pick $(Y_1, Y_2, \ldots, Y_N)$ independently from $P_{\text{cond}}(\cdot|\Xi_i)$. Note that ultimately Z_i is a measurable function ϕ of Ξ_i and Y_i, namely $Z_i = \phi(\Xi_i, Y_i)$.

Theorem 5 *For hierarchical sampling with finite exchangeability, we have*

$$\mathbb{E}_{\hat{P}}[\Delta_{\epsilon-\text{Gen}}(\hat{\theta}(\epsilon))] = \epsilon^2 \Bigg[\frac{1}{N}\mathbb{E}_{Z\sim P}\big[\mathcal{K}_\lambda(Z, Z)\big] - \frac{N+n-1}{Nn}\mathbb{E}_{(Z,Z')\sim P^{\otimes 2}}\big[\mathcal{K}_\lambda(Z, Z')\big]$$
$$+ \frac{N-1}{Nn}\mathbb{E}_{X\sim P,(Y,Y')\sim P_{\text{cond}}(\cdot|X)^{\otimes 2}}\big[\mathcal{K}_\lambda(\phi(X, Y), \phi(X, Y'))\big]\Bigg] + o(\epsilon^2). \quad (3)$$

The diagonal term measures self-influence, while the same-parent term captures correlations between different views of the same latent sample. This provides a simple mechanism by which augmentations and hierarchical sampling modify generalization.

For square-loss mean estimators, the second-order expansion is exact (see Appendix J).

5. Experiments

We compare empirical generalization gaps with influence-based predictions in high-dimensional linear regression, for both i.i.d. and hierarchical sampling settings. When we train on a particular dataset, we can estimate test loss and train loss by splitting the dataset into test and train sets. Suppose we have N training data points and N' test data points. To estimate the influence-based prediction, we may use either training or test samples. Let the optimal parameter be $\hat{\theta}_N$.

To explain the idea, we will focus on the case of $\mathbb{E}_{\hat{P}}[\Delta_{\epsilon-\mathrm{Gen}}(\hat{\theta}(\epsilon))]$ with $\lambda = 0$, in the i.i.d. sample case. Taking the leading-order answer in ϵ and setting $\epsilon = 1$, we get $\frac{1}{N}\mathbb{E}_{Z\sim P}\big[\mathcal{K}_0(Z,Z)\big]$, since $\mathbb{E}_{(Z,Z')\sim P^{\otimes 2}}\big[\mathcal{K}_0(Z,Z')\big] = 0$.

Let

$$\hat{H}_0 := \frac{1}{N}\sum_{z_{\mathrm{train}}} \nabla_\theta^2 l(z_{\mathrm{train}}, \hat{\theta}_N), \text{ and, let } \hat{K}_0(z,z') := \nabla_\theta l(z,\hat{\theta}_N)^\top \hat{H}_0^{-1} \nabla_\theta l(z',\hat{\theta}_N).$$

Then we also have two estimates:

$$\hat{\Delta}_{\mathrm{gen-test}} := \frac{1}{N'}\sum_{z_{\mathrm{test}}} \hat{K}_0(z_{\mathrm{test}}, z_{\mathrm{test}}), \quad \hat{\Delta}_{\mathrm{gen-train}} := \frac{1}{N}\sum_{z_{\mathrm{train}}} \hat{K}_0(z_{\mathrm{train}}, z_{\mathrm{train}}).$$

The two estimates are close in well-generalizing regimes, while the test-based estimate becomes increasingly conservative near memorization.

Unless otherwise stated, all experiments use an 80/20 train-test split for 1000 data points. We report the empirical generalization gap, which is average test loss minus average train loss and compare it to the corresponding diagonal influence prediction obtained from the explicit formula for the loss $\frac{1}{2}\sum_{i=1}^{N}(y_i - \theta^\top x_i)^2$.

5.1. Linear Regression with Independent Samples

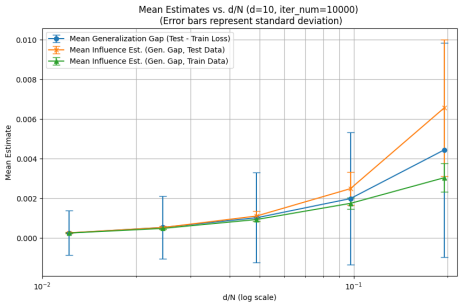

Figure 1: Comparison of train-based and test-based influence-function estimates of the expected generalization gap across learning regimes. In well-generalizing regimes, the two estimates closely agree. As generalization worsens and the model approaches memorization, the test-based estimate becomes systematically more conservative, suggesting increasing concentration of self-influence and the growing importance of higher-order corrections beyond the leading-order von Mises expansion.

We consider teacher–student Gaussian linear regression:

$$x_i \sim \mathcal{N}(0, I_d), \qquad y_i = \theta^{*\top} x_i + \epsilon_i,$$

with $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$. Fig. 1 compares empirical generalization gaps with influence-based predictions over 10,000 trials. The train-based and test-based estimators bracket the true gap and separate near the memorization regime $d/N \to 1^-$.

5.2. Linear Regression with Hierarchical Samples

For hierarchical sampling, with replacement draws from n latent samples $x_\alpha \sim \mathcal{N}(0, I_d)$ generate noisy views

$$\tilde{x}_i = x_\alpha + \zeta_i.$$

with $\zeta_i \sim \mathcal{N}(0, \tau^2), \epsilon_i \sim \mathcal{N}(0, \sigma^2)$ and $y_i = \theta^{*\top} x_\alpha + \epsilon_i, i = 1, \ldots, N_{\text{total}}$. We report results over 100 trials.

Fig. 2 contrasts the i.i.d. (panel (a)) and hierarchical (panel (b)) distributions of the empirical generalization gap and the corresponding influence-based estimates. We use the training-based estimator since, for $d = 10$ and $N = 1000$, the system remains in the well-generalizing regime.

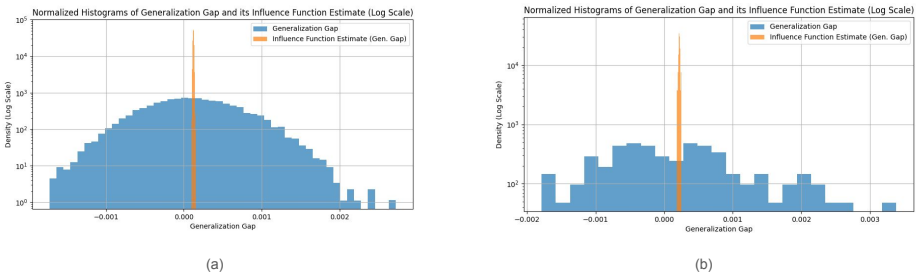

Figure 2: Normalized histograms of the empirical generalization gap (blue) and the influence function estimates of its expectation (orange). (a) 10^4 trials of i.i.d samples, (b) 100 trials of hierarchical samples.

6. Conclusions

We showed that expected generalization gaps are governed by self-influence and same-parent cross-influence. Influence-based estimators accurately track generalization in well-generalizing regimes and become conservative near memorization. These results suggest that influence geometry is a useful framework for understanding generalization in modern learning systems.

Limitations: Our approach of expanding in the parameter ϵ shows fluctuations as well as biases together, but the results may be obtained separately from large sample asymptotics (see Appendix H). A question remains whether all the results from this approach can be obtained by taking some appropriate large sample limit. Another possible limitation associated with this approach is the assumption that parameters change smoothly as we turn on the effect of finite samples. This is not true for problems in the underconstrained regime without strong regularization terms. This may thus inhibit a direct attack on problems like double descent [3].

References

[1] Hirotugu Akaike. Akaike's information criterion. In *International encyclopedia of statistical science*, pages 41–42. Springer, 2025.

[2] Randall Balestriero, Mark Ibrahim, Vlad Sobal, Ari Morcos, Shashank Shekhar, Tom Goldstein, Florian Bordes, Adrien Bardes, Gregoire Mialon, Yuandong Tian, et al. A cookbook of self-supervised learning. *arXiv preprint arXiv:2304.12210*, 2023.

[3] Mikhail Belkin, Daniel Hsu, Siyuan Ma, and Soumik Mandal. Reconciling modern machine-learning practice and the classical bias–variance trade-off. *Proceedings of the National Academy of Sciences*, 116(32):15849–15854, 2019.

[4] Persi Diaconis and David Freedman. Finite exchangeable sequences. *The Annals of Probability*, pages 745–764, 1980.

[5] Luisa Turrin Fernholz. *Von Mises calculus for statistical functionals*, volume 19. Springer Science & Business Media, 2012.

[6] Frank R Hampel. The influence curve and its role in robust estimation. *Journal of the american statistical association*, 69(346):383–393, 1974.

[7] Moritz Hardt and Benjamin Recht. *Patterns, predictions, and actions: Foundations of machine learning*. Princeton University Press, 2022.

[8] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

[9] Pang Wei Koh and Percy Liang. Understanding black-box predictions via influence functions. In *International conference on machine learning*, pages 1885–1894. PMLR, 2017.

[10] Ritei Shibata. Statistical aspects of model selection. In *From data to model*, pages 215–240. Springer, 1989.

[11] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International conference on machine learning*, pages 2256–2265. pmlr, 2015.

[12] Yang Song and Stefano Ermon. Generative modeling by estimating gradients of the data distribution. *Advances in neural information processing systems*, 32, 2019.

[13] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*, 2020.

[14] Koji Takeuchi. Distribution of informational statistics and a criterion of model fitting. suri-kagaku (mathematical sciences) 153 12-18, 1976.

[15] Aad W Van der Vaart. *Asymptotic statistics*, volume 3. Cambridge university press, 2000.

[16] Halbert White. Maximum likelihood estimation of misspecified models. *Econometrica: Journal of the econometric society*, pages 1–25, 1982.

Appendix A. Examples

Example: Maximum Likelihood Estimation. Let $\mathcal{L}(\theta, z)$ be the (potentially misspecified) likelihood for the data point z. We can define negative log-likelihood $l(z, \theta) := -\ln \mathcal{L}(\theta, z)$. With this definition, $\hat{\theta} := \arg\min_{\theta \in \Theta} L(\theta, \hat{P})$ is the maximum likelihood estimator.

Example: Supervised learning with Square Loss. Let z denote the pair (x, y) where $x \in \mathbb{R}^{d_{\text{in}}}$, $y \in \mathbb{R}^{d_{\text{out}}}$. We would typically have a predictor function $f(\cdot, \theta) : \mathbb{R}^{d_{\text{in}}} \to \mathbb{R}^{d_{\text{out}}}$ parameterized by the parameters $\theta \in \Theta$. We also need a loss function $l : \mathbb{R}^{d_{\text{in}}+d_{\text{out}}} \times \Theta \to \mathbb{R}$, concretely, $l(z, \theta) = \frac{1}{2} \|y - f(x, \theta)\|^2$, $\|\cdot\|$ being the l_2 norm, for regression.

Appendix B. The Four Losses and the Optimization Landscape

The four losses of interest are organized in Table 1.

Table 1: The four combinations for the loss functional.

DATA DISTRN. \ OPTIMAL PARAM.	θ^*_λ	$\hat{\theta}_\lambda$
P	$L(\theta^*_\lambda, P)$	$L(\hat{\theta}_\lambda, P)$
$\hat{P}$	$L(\theta^*_\lambda, \hat{P})$	$L(\hat{\theta}_\lambda, \hat{P})$

The nature of the differences between various losses and the ϵ-interpolation is represented in Fig. 3.

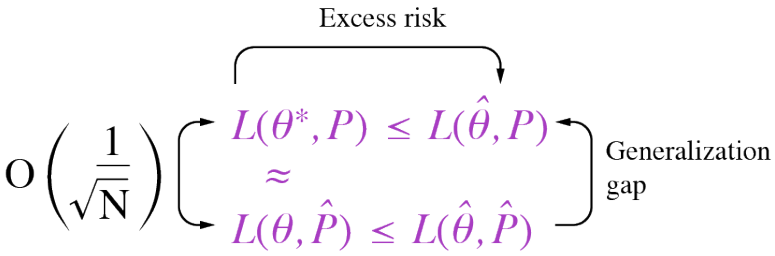

Figure 3: The relation between different losses for $\lambda = 0$

Here is the objective function (=loss + regularization) landscape as a function of θ and ϵ in Fig. 4.

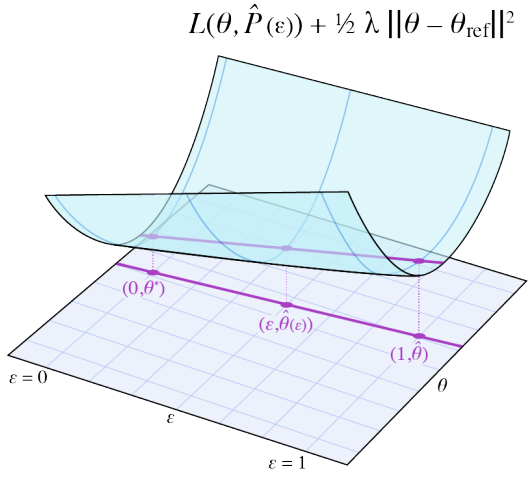

Figure 4: The relation between different losses for $\lambda = 0$.

Appendix C. Further Preliminaries

C.1. Influence Function

Following Koh and Liang [9], the formula describing how the loss for a specific test point z_{test} changes, owing to an upweighting of training point z is given by the influence function

$$\mathcal{I}_{\text{up,loss}}(z, z_{\text{test}}) := \left.\frac{dl(z_{\text{test}}; \hat{\theta}_{\epsilon,z})}{d\epsilon}\right|_{\epsilon=0} = -\nabla_\theta l(z_{\text{test}}; \hat{\theta}_\lambda)^\top (H_{\hat{\theta}} + \lambda \mathbb{I})^{-1} \nabla_\theta l(z; \hat{\theta}_\lambda). \tag{4}$$

The Hessian $H_{\hat{\theta}}$ is defined as the average of the second derivatives of the training loss across all n points, namely, $H_{\hat{\theta}} := \frac{1}{n}\sum_{i=1}^n \nabla_\theta^2 l(z_i, \hat{\theta}_\lambda)$.

Appendix D. Influence Kernel

We first make a few comments about members of the algebraic tensor product $L^1(P) \otimes L^1(P)$, which are functions $\psi : \mathbb{R}^D \times \mathbb{R}^D \to \mathbb{R}$ of the form

$$\psi(z, z') = \sum_{i=1}^r a_i f_i(z) g_i(z'),$$

where $f_i, g_i \in L^1(P)$ for all $i = 1, \ldots, r$. Essentially, these functions of (z, z') are finite linear combinations of products of two integrable functions, one of z and the other of z'.

Remark 6 *Under Assumption 13,* $\mathcal{K}_\lambda, \tilde{\mathcal{K}}_\lambda \in L^1(P) \otimes L^1(P)$.

Lemma 7 *If* $\psi \in L^1(P) \otimes L^1(P)$,

$$\int \psi(z, z') d(P \times P)(z, z') = \int \Big(\int \psi(z, z') dP(z) \Big) dP(z') = \int \Big(\int \psi(z, z') dP(z') \Big) dP(z).$$

Namely, we can do iterated integrals for the double integrals.

Proof Since $|\psi(z,z')| \leq \sum_{i=1}^{r} |a_i||f_i(z)||g_i(z')|$,

$$\int |\psi| d(P \times P) \leq \sum_{i=1}^{r} |a_i| \int |f_i||g_i| d(P \times P) = \sum_{i=1}^{r} |a_i| \int |f_i| dP \int |g_i| dP < \infty,$$

where we have used Tonelli's theorem in the intermediate step. Therefore, $\psi \in L^1(P \times P)$, Fubini's theorem applies. ■

Remark 8 *For $\psi \in L^1(P) \otimes L^1(P)$, Fubini's theorem could be applied to its integral of the product of the signed measure $\int \psi d((\hat{P} - P) \times (\hat{P} - P))$. This because, using very similar arguments as for the Lemma 7,*

$$\int |\psi| d((\hat{P} + P) \times (\hat{P} + P)) \leq \sum_{i=1}^{r} |a_i| \int |f_i| d(\hat{P} + P) \int |g_i| d(\hat{P} + P) < \infty.$$

Remark 9 *For $\psi \in L^1(P) \otimes L^1(P)$, if λ, ν are signed measures formed by combining $\hat{P}, P$, we often write the integral of ψ under the measure $\lambda \times \nu$ as $\int \psi(z,z') d\lambda(z) d\nu(z')$ rather than as $\int \psi(z,z') d(\lambda \times \nu)(z,z')$, since Fubini's theorem could be applied to the integral,.*

Appendix E. Assumptions

Assumption 10 *Both $L(\theta, P)$ and $L(\theta, \hat{P})$ are C^2 in θ, in an open neighborhood U of θ^*_λ. We assume that $L(\theta, \hat{P}(\epsilon)) + \frac{1}{2}\lambda \|\theta - \theta_{\text{ref}}\|^2$ has a unique minimum at $\hat{\theta}_\lambda(\epsilon)$ in U for $\epsilon \in [0,1]$ with the Hessian being positive definite at that unique minimum. We further assume that both $L(\theta, P)$ and $L(\theta, \hat{P})$ are second differentiable in U.*

Here we make some strong assumptions about loss function statistics and its Taylor expansion in ϵ.

Assumption 11 *Consider the function $F(\epsilon, \epsilon') = L(\hat{\theta}(\epsilon), \hat{P}(\epsilon'))$. We define the function on $I \times I$, where $I \supseteq [0,1]$ is an open interval. We assume F is C^2 in a neighborhood of $(0,0)$, allowing a second order Taylor polynomial $\mathcal{P}_2$ of F at $(0,0)$. Let the remainder be*

$$R_2(\epsilon, \epsilon') := F(\epsilon, \epsilon') - \mathcal{P}_2(\epsilon, \epsilon').$$

We assume that $\mathbb{E}_{\hat{P}}[|R_2(\epsilon, \epsilon')|] = o((\max(\epsilon, \epsilon'))^2)$.

Remark 12 *Note that, for a random variable V, if $\mathbb{E}[|V|] = o(a)$ as $a \to 0^+$, V is $o_P(a)$, by Markov's inequality.*

Assumption 13 *We assume $\nabla_\theta l(z, \theta)$ exists for all $z \in \mathbb{R}^D, \theta \in U$, and for $\theta \in U$, $\nabla_\theta l(\cdot, \theta)$ is measurable, with the property that $||\nabla_\theta l(z, \theta)|| \leq g(z), z \in \mathbb{R}^D$, for some measurable g with $\mathbb{E}_{Z \sim P}[g(Z)] < \infty$.*

One particular property of the influence kernel is going to be quite useful for us.

Proposition 14 *If Assumption 13 is true, then*

$$\mathbb{E}_{(Z,Z')\sim P^{\otimes 2}}\big[\mathcal{K}_\lambda(Z,Z')\big] := \int \mathcal{K}_\lambda(z,z')dP(z)dP(z') = \lambda^2(\theta^*_\lambda - \theta_{\text{ref}})^\top (H^* + \lambda\mathbb{I})^{-1}(\theta^*_\lambda - \theta_{\text{ref}}) \tag{5}$$

Proof The Lemma 7 allows us to do the double integral as an iterated integral. $\mathbb{E}_{Z\sim P}\big[\|\nabla_\theta l(Z,\theta^*)\|\big]$ being finite allows us to pull the gradient from inside the single integral to the outside for each component, since $|\frac{\partial l(z,\theta^*)}{\partial \theta_\alpha}| \le \|\nabla_\theta l(z,\theta^*)\|$, for each α.

$$\mathbb{E}_{(Z,Z')\sim P^{\otimes 2}}\big[\mathcal{K}_\lambda(Z,Z')\big] = \int \mathcal{K}_\lambda(z,z')dP(z)dP(z') = \int \nabla_\theta l(z,\theta^*)^\top (H^* + \lambda\mathbb{I})^{-1}\nabla_\theta l(z',\theta^*)dP(z)dP(z')$$
$$= \int \nabla_\theta l(z,\theta^*)^\top (H^* + \lambda\mathbb{I})^{-1}\Big(\int \nabla_\theta l(z',\theta^*)dP(z')\Big)dP(z) = \int \nabla_\theta l(z,\theta^*)^\top (H^* + \lambda\mathbb{I})^{-1}\nabla_\theta\Big(\int l(z,\theta^*)dP(z')\Big)d$$
$$= \int \nabla_\theta l(z,\theta^*)^\top (H^* + \lambda\mathbb{I})^{-1}\nabla_\theta L(\theta^*,P)dP(z) = \nabla_\theta L(\theta^*,P)^\top (H^* + \lambda\mathbb{I})^{-1}\nabla_\theta L(\theta^*,P)$$
$$= \lambda^2(\theta^* - \theta_{\text{ref}})^\top (H^* + \lambda\mathbb{I})^{-1}(\theta^* - \theta_{\text{ref}}),$$

since $\nabla_\theta L(\theta^*,P) = -\lambda(\theta^* - \theta_{\text{ref}})$. ■

So, if $\lambda = 0$, then $\mathbb{E}_{(Z,Z')\sim P^{\otimes 2}}\big[\mathcal{K}_\lambda(Z,Z')\big] = 0$.

Appendix F. Proof of the Generalization Gap Results

F.1. Proof of Proposition 3

Proof Define the optimization objective function gradient vector

$$G(\epsilon,\theta) := \nabla_\theta L(\theta,\hat{P}(\epsilon)) + \lambda(\theta - \theta_{\text{ref}}). \tag{6}$$

The optimality condition is $G(\epsilon,\hat{\theta}(\epsilon)) = 0$. We will track the evolution of the optimum $\hat{\theta}(\epsilon)$ in the θ^* neighborhood, as one changes ϵ. Since under Assumption 10, $G(\epsilon,\theta)$ is C^1, and $\nabla_\theta G(\epsilon,\theta)$ is positive definite. So $\nabla_\theta G(\epsilon,\theta)$, which is also the Hessian, is invertible at $\theta = \hat{\theta}(\epsilon)$. The implicit function theorem says that the function $\hat{\theta}(\cdot)$ is differentiable and satisfies

$$\nabla_\theta G(\epsilon,\theta)\frac{d\hat{\theta}(\epsilon)}{d\epsilon} + \frac{\partial}{\partial\epsilon}G(\epsilon,\theta) = 0. \tag{7}$$

Plugging $L(\theta,\hat{P}(\epsilon)) = L(\theta,P) + \epsilon L(\theta,\hat{P} - P)$ into Eq. (6) for $G(\epsilon,\theta)$, we find

$$G(\epsilon,\theta) = \nabla_\theta L(\theta,P) + \lambda(\theta - \theta_{\text{ref}}) + \epsilon\nabla_\theta L(\theta,\hat{P} - P). \tag{8}$$

So, Eq. (7) becomes

$$\big(\nabla^2_\theta L(\theta,P) + \lambda\mathbb{I} + \epsilon\nabla^2_\theta L(\theta,\hat{P} - P)\big)\frac{d\hat{\theta}(\epsilon)}{d\epsilon} + \nabla_\theta L(\theta,\hat{P} - P) = 0. \tag{9}$$

Thus,

$$\frac{d\hat{\theta}(\epsilon)}{d\epsilon} = -\big(\nabla^2_\theta L(\theta,P) + \lambda\mathbb{I} + \epsilon\nabla^2_\theta L(\theta,\hat{P} - P)\big)^{-1}\nabla_\theta L(\theta,\hat{P} - P). \tag{10}$$

If we let $\theta \to \theta^*$ and $\epsilon \to 0$, we find

$$\begin{aligned}\left.\frac{d\hat{\theta}(\epsilon)}{d\epsilon}\right|_{\epsilon=0} &= -\big(\nabla_\theta^2 L(\theta^*, P) + \lambda \mathbb{I}\big)^{-1} \nabla_\theta L(\theta^*, \hat{P} - P) \\ &= -H_\lambda^{-1} \int \nabla_\theta l(z, \theta^*) d(\hat{P} - P)(z).\end{aligned} \tag{11}$$

We have used $L(\theta^*, \hat{P} - P) = \int l(z, \theta^*) d(\hat{P} - P)(z)$ following Eq. (**??**), and denoted the Hessian as $H_\lambda := \nabla_\theta^2 L(\theta^*, P) + \lambda \mathbb{I}$.

We wish to understand how much $L(\hat{\theta}(\epsilon), P)$ differs from $L(\hat{\theta}(\epsilon), \hat{P}(\epsilon))$ on average. Notice that $L(\hat{\theta}(1), \hat{P}(1))$ is the training loss and $L(\hat{\theta}(1), P)$ is the test loss, so $L(\hat{\theta}(\epsilon), P) - L(\hat{\theta}(\epsilon), \hat{P}(\epsilon))$ is the generalization gap when $\epsilon = 1$. Using Definition 2 of ϵ-generalization gap: $\Delta_{\epsilon-\mathrm{Gen}}(\hat{\theta}(\epsilon)) := L(\hat{\theta}(\epsilon), P) - L(\hat{\theta}(\epsilon), \hat{P}(\epsilon))$, and, combining the two loss functions, we obtain

$$\Delta_{\epsilon-\mathrm{Gen}}(\hat{\theta}(\epsilon)) = L(\hat{\theta}(\epsilon), P - \hat{P}(\epsilon)) = -\epsilon L(\hat{\theta}(\epsilon), \hat{P} - P). \tag{12}$$

We wish to Taylor-expand the ϵ-generalization gap around $\epsilon = 0$. For this, we first compute

$$\begin{aligned}\frac{d}{d\epsilon}\Delta_{\epsilon-\mathrm{Gen}}(\hat{\theta}(\epsilon)) = \frac{d}{d\epsilon}[L(\hat{\theta}(\epsilon), P - \hat{P}(\epsilon))] &= -\frac{d}{d\epsilon}\big(\epsilon L(\hat{\theta}(\epsilon), \hat{P} - P)\big) \\ &= -L(\hat{\theta}(\epsilon), \hat{P} - P) - \epsilon[\nabla_\theta L(\hat{\theta}(\epsilon), \hat{P} - P)]^\top \frac{d\hat{\theta}(\epsilon)}{d\epsilon},\end{aligned} \tag{13}$$

then

$$\begin{aligned}\frac{d^2}{d\epsilon^2}\Delta_{\epsilon-\mathrm{Gen}}(\hat{\theta}(\epsilon)) &= \frac{d}{d\epsilon}\Big[\frac{d}{d\epsilon}\Delta_{\epsilon-\mathrm{Gen}}(\hat{\theta}(\epsilon))\Big] \\ &= \frac{d}{d\epsilon}\Big[- L(\hat{\theta}(\epsilon), \hat{P} - P) - \epsilon[\nabla_\theta L(\hat{\theta}(\epsilon), \hat{P} - P)]^\top \frac{d\hat{\theta}(\epsilon)}{d\hat{\epsilon}}\Big] \\ &= -[\nabla_\theta L(\hat{\theta}(\epsilon), \hat{P} - P)]^\top \frac{d\hat{\theta}(\epsilon)}{d\epsilon} - [\nabla_\theta L(\hat{\theta}(\epsilon), \hat{P} - P)]^\top \frac{d\hat{\theta}(\epsilon)}{d\epsilon} \\ &\quad - \epsilon \frac{d\hat{\theta}(\epsilon)}{d\epsilon}^\top \nabla_\theta^2 L(\hat{\theta}(\epsilon), \hat{P} - P)) \frac{d\hat{\theta}(\epsilon)}{d\epsilon} - \epsilon[\nabla_\theta L(\hat{\theta}(\epsilon), \hat{P} - P)]^\top \frac{d^2\hat{\theta}(\epsilon)}{d\epsilon^2} \\ &= -2[\nabla_\theta L(\hat{\theta}(\epsilon), \hat{P} - P)]^\top \frac{d\hat{\theta}(\epsilon)}{d\epsilon} - \epsilon \frac{d\hat{\theta}(\epsilon)}{d\epsilon}^\top \nabla_\theta^2 L(\hat{\theta}(\epsilon), \hat{P} - P)) \frac{d\hat{\theta}(\epsilon)}{d\epsilon} - \epsilon[\nabla_\theta L(\hat{\theta}(\epsilon), \hat{P} - P)]^\top \frac{d^2}{d} \\ &= -2[\nabla_\theta L(\hat{\theta}(\epsilon), \hat{P} - P)]^\top \frac{d\hat{\theta}(\epsilon)}{d\epsilon} + O_P(\epsilon).\end{aligned} \tag{14}$$

Using these two expressions, we can write down a Taylor expansion in ϵ for $L(\theta(\epsilon), P) - L(\theta(\epsilon), \hat{P}(\epsilon))$ at $\theta = \theta^*$:

$$
\begin{aligned}
L(\hat{\theta}(\epsilon), P) - L(\hat{\theta}(\epsilon), \hat{P}(\epsilon)) &= \frac{d}{d\epsilon}\big[L(\hat{\theta}(\epsilon), P) - L(\hat{\theta}(\epsilon), \hat{P}(\epsilon))\big]\Big|_{\epsilon=0}\epsilon + \frac{1}{2}\frac{d^2}{d\epsilon^2}\big[L(\hat{\theta}(\epsilon), P) - L(\hat{\theta}(\epsilon), \hat{P}(\epsilon))\big]\Big|_{\epsilon=0}\epsilon^2 + o_P \\
&= \left[- L(\hat{\theta}(\epsilon), \hat{P} - P) - \epsilon[\nabla_\theta L(\hat{\theta}(\epsilon), \hat{P} - P)]^\top \frac{d\hat{\theta}(\epsilon)}{d\epsilon}\right]\Big|_{\epsilon=0}\epsilon + \frac{1}{2}\left[- 2\nabla_\theta L(\hat{\theta}(\epsilon), \hat{P} - P)]^\top \frac{d\hat{\theta}(\epsilon)}{d\epsilon} + O(\epsilon)\right]\Big|_{\epsilon=0}\epsilon^2 + o \\
&= -L(\theta^*, \hat{P} - P)\epsilon - [\nabla_\theta L(\theta^*, \hat{P} - P)]^\top \frac{d\hat{\theta}(\epsilon)}{d\epsilon}\Big|_{\epsilon=0}\epsilon^2 + o_P(\epsilon^2).
\end{aligned}
\tag{15}
$$

Then, using Eq. (11), we find

$$
\begin{aligned}
\Delta_{\epsilon-\mathrm{Gen}}(\hat{\theta}(\epsilon)) &:= L(\hat{\theta}(\epsilon), P) - L(\hat{\theta}(\epsilon), \hat{P}(\epsilon)) \\
&= -L(\theta^*, \hat{P} - P)\epsilon + [\nabla_\theta L(\theta^*, \hat{P} - P)]^\top \left(\nabla^2_\theta L(\theta^*, P) + \lambda \mathbb{I}\right)^{-1} \nabla_\theta L(\theta^*, \hat{P} - P)\epsilon^2 + o(\epsilon^2) \\
&= -L(\theta^*, \hat{P} - P)\epsilon + [\nabla_\theta L(\theta^*, \hat{P} - P)]^\top H_\lambda^{-1} \nabla_\theta L(\theta^*, \hat{P} - P)\epsilon^2 + o_P(\epsilon^2) \\
&= -L(\theta^*, \hat{P} - P)\epsilon + \int \nabla_\theta l(z, \theta^*)^\top H_\lambda^{-1} \nabla_\theta l(z', \theta^*) d(\hat{P} - P)(z) d(\hat{P} - P)(z') + o_P(\epsilon^2) \\
&= -L(\theta^*, \hat{P} - P)\epsilon + \int \mathcal{K}_\lambda(z, z') d(\hat{P} - P)(z) d(\hat{P} - P)(z') + o_P(\epsilon^2),
\end{aligned}
\tag{16}
$$

using Definition 1 for the influence kernel $\mathcal{K}_\lambda$.

A comment on the $o_P(\epsilon^2)$ term coming from the difference of remainders. The expectations of the absolute value of these remainders are $o(\epsilon^2)$, by Assumption 11. This would be true of the difference as well. As a result, it is $o_P(\epsilon^2)$. ■

When we compute the expected value of the ϵ-generalization gap, the first term contributes zero and we get the following corollary.

Corollary 15 *Under Assumptions 10 and 11,*

$$
\mathbb{E}_{\hat{P}}[\Delta_{\epsilon-\mathrm{Gen}}(\hat{\theta}(\epsilon))] = \epsilon^2 \mathbb{E}_{\hat{P}}\left[\int \mathcal{K}_\lambda(z, z') d(\hat{P} - P)(z) d(\hat{P} - P)(z') \right] + o(\epsilon^2). \tag{17}
$$

Appendix G. Taking the Expectation

Both expected ϵ-generalization gap and (expected) excess risk have leading order contributions in an ϵ-power expansion with expressions of the form $\mathbb{E}_{\hat{P}}\left[\int \psi(z, z') d(\hat{P} - P)(z) d(\hat{P} - P)(z') \right]$, where $\psi \in L^1(P) \otimes L^1(P)$, the algebraic tensor product, and $\hat{P}$ is an empirical sample distribution.

Proposition 16 *If $\psi \in L^1(P) \otimes L^1(P)$,*

$$
\mathbb{E}_{\hat{P}}\left[\int \psi(z, z') d(\hat{P} - P)(z) d(\hat{P} - P)(z') \right] = \mathbb{E}_{\hat{P}}\left[\int \psi(z, z') d\hat{P}(z) d\hat{P}(z') \right] - \mathbb{E}_{Z, Z' \sim P^{\otimes 2}}\left[\psi(Z, Z') \right]. \tag{18}
$$

Often, for learning problems, one considers n i.i.d. samples from P. This is appropriate for many unsupervised and supervised learning problems. However, in the self-supervised setting [2], we need a two-stage sampling strategy: n i.i.d. original samples, and the N augmentations/noise corruptions, based on these original samples. Such hierarchical samples include 'correlated' ones, which are different corruptions/augmentations of the same original sample. This is true of diffusion models [8, 11–13] as well. Hence, we start by assuming something weaker, namely, finite exchangeability [4]. We get the simple proposition:

Proposition 17 *If $(Z_1, Z_2, \ldots, Z_N)$ is an exchangeable finite sequence with $\hat{P} = \frac{1}{N}\sum_{i=1}^{N} \delta_{Z_i}$, then, for any measurable $\psi : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ with $\mathbb{E}\big[\psi(Z_i, Z_j)\big]$ well-defined for any $1 \leq i, j \leq N$,*

$$\begin{aligned}\mathbb{E}_{\hat{P}}\left[\int \psi(z,z')d\hat{P}(z)d\hat{P}(z')\right] &= \frac{1}{N^2}\mathbb{E}\left[\sum_{i,j}\psi(Z_i,Z_j)\right] \\ &= \frac{1}{N}\mathbb{E}\big[\psi(Z_1,Z_1)\big] + \frac{N-1}{N}\mathbb{E}\big[\psi(Z_1,Z_2)\big] \end{aligned} \tag{19}$$

If $(Z_1, Z_2, \ldots, Z_N)$ are i.i.d then

$$\mathbb{E}\big[\psi(Z_1,Z_1)\big] = \mathbb{E}_{Z\sim P}\big[\psi(Z,Z)\big], \qquad \mathbb{E}\big[\psi(Z_1,Z_2)\big] = \mathbb{E}_{(Z,Z')\sim P^{\otimes 2}}\big[\psi(Z,Z')\big] \tag{20}$$

The consequence of Propositions 16&17 is the following.

Proposition 18 *If $(Z_1, Z_2, \ldots, Z_N)$ are i.i.d then*

$$\mathbb{E}_{\hat{P}}\left[\int \psi(z,z')d(\hat{P}-P)(z)d(\hat{P}-P)(z')\right] = \frac{1}{N}\Big[\mathbb{E}_{Z\sim P}\big[\psi(Z,Z)\big] - \mathbb{E}_{(Z,Z')\sim P^{\otimes 2}}\big[\psi(Z,Z')\big]\Big] \tag{21}$$

However, suppose we have the following hierarchical sampling setting. Sample $(X_1, X_2, \ldots, X_n)$ independently from P, and then sample $(\Xi_1, \Xi_2, \ldots, \Xi_N)$ independently from $(\frac{1}{n}\sum_{\alpha=1}^{n}\delta_{X_\alpha})$. Then pick $(Y_1, Y_2, \ldots, Y_N)$ independently from $P_{\text{cond}}(\cdot|\Xi_i)$. Note that ultimately Z_i is a measurable function ϕ of Ξ_i and Y_i, namely $Z_i = \phi(\Xi_i, Y_i)$. Application of these observations gives:

Proposition 19 *With these sampling rules, we obtain the following result.*

$$\begin{aligned}\mathbb{E}_{\hat{P}}\left[\int \psi(z,z')d(\hat{P}-P)(z)d(\hat{P}-P)(z')\right] &= \frac{1}{N}\mathbb{E}_{Z\sim P}\big[\psi(Z,Z)\big] - \frac{N+n-1}{Nn}\mathbb{E}_{(Z,Z')\sim P^{\otimes 2}}\big[\psi(Z,Z')\big] \\ &+ \frac{N-1}{Nn}\mathbb{E}_{X\sim P,(Y,Y')\sim P_{\text{cond}}(\cdot|X)^{\otimes 2}}\big[\psi(\phi(X,Y),\phi(X,Y'))\big] \end{aligned} \tag{22}$$

Note that if we take the limit $n \to \infty$ on the RHS of Eq. (22), we get Eq. (21), as expected.

Combining propositions 18 and 19 with Eqs. 17, we get theorems 4 and 5, which are our main results.

Appendix H. Relationship to Asymptotic Analysis

Let us start by asking how our results relate to the traditional large sample limit results [15], which correspond to large N but $\epsilon = 1$. To illustrate this relationship, we begin with the case of ϵ-generalization gap, when the samples are i.i.d. For clarity, the the finite sample optimal parameter would be indicated as $\hat{\theta}_N$ rather than as $\hat{\theta}$. Proposition 3 has the first two terms of a Taylor series in ϵ. The leading term is related to the following asymptotic result about the variance of the generalization gap at the sample-dependent optimal parameter. For simplicity, we have $\lambda = 0$.

Proposition 20 *When Z_i are i.i.d. and $\mathbb{E}_{Z\sim P}\left[l(Z,\theta^*)^2\right] < \infty$*

$$\lim_{N\to\infty} N\mathrm{Var}_{(Z_1,\ldots,Z_N)\sim P^{\otimes N}}\left(\Delta_{\mathrm{gen}}(\hat{\theta}_N)\right) = \mathrm{Var}_{Z\sim P}(l(Z,\theta^*)).$$

The second term, on the other hand, is related to the asymptotic properties of the expectation of generalization gap

Proposition 21 *When Z_i are i.i.d. and $\mathbb{E}_{Z\sim P}\left[\mathcal{K}_0(Z,Z)\right]$ is well-defined*

$$\lim_{N\to\infty} N\mathbb{E}_{(Z_1,\ldots,Z_N)\sim P^{\otimes N}}\left[\Delta_{\mathrm{gen}}(\hat{\theta}_N)\right] = \mathbb{E}_{Z\sim P}\left[\mathcal{K}_0(Z,Z)\right].$$

This proposition is related to the results of Akaike [1], Shibata [10], Takeuchi [14], White [16] where the expected log-likelihood bias of a mis-specified model: $\mathbb{E}[\ell_n(\hat{\theta}) - nP\ell_{\hat{\theta}}] \to -\frac{1}{2}\mathrm{tr}(I^{-1}J)$, with $I = -P\ddot{\ell}_{\theta_0}$, $J = \mathrm{Var}_P(\dot{\ell}_{\theta_0})$, in typical notation of [15].

For the proofs of these two propositions, see Appendix I.

Our results suggest that, for the hierarchical sampling case, there should be asymptotic results analogous to Propositions 20, and 21 as $N, n \to \infty$ with N/n remaining fixed. We are not aware of any such results in the literature, though.

Appendix I. Proofs of the Asymptotic Relationships

Here, we provide a bridge between our approach and that of the asymptotics literature, establishing correspondence between notation and concepts.

Let

$$P_N := \frac{1}{N}\sum_{i=1}^{N}\delta_{Z_i},$$

where $Z_1, \ldots, Z_N$ are i.i.d. samples from the population distribution P.We define

$$L(\theta, P) := P\ell_\theta := \int \ell(z,\theta)\, dP(z),$$

and assume that

$$\theta^* = \arg\min_\theta P\ell_\theta.$$

Hence

$$P\dot{\ell}_{\theta^*} = 0,$$

where

$$\dot{\ell}_{\theta^*}(z) := \nabla_\theta \ell(z,\theta^*).$$

We also define the population Hessian

$$H := \nabla_\theta^2 L(\theta^*, P).$$

Assume:

1. $\hat{\theta}_N \xrightarrow{P} \theta^*$,
2. $\ell(z, \theta)$ is twice continuously differentiable in a neighborhood of θ^*,
3. H is invertible,
4. the standard stochastic equicontinuity assumptions for smooth M-estimators hold.

Under these assumptions, the classical M-estimator expansion gives

$$\sqrt{N}(\hat{\theta}_N - \theta^*) = -H^{-1}\sqrt{N}(P_N - P)\dot{\ell}_{\theta^*} + o_P(1).$$

Equivalently,

$$\hat{\theta}_N - \theta^* = -H^{-1}(P_N - P)\dot{\ell}_{\theta^*} + o_P(N^{-1/2}). \tag{23}$$

We define the influence kernel

$$K_0(z, z') = \dot{\ell}_{\theta^*}(z)^\top H^{-1} \dot{\ell}_{\theta^*}(z').$$

I.1. Proof of Proposition 20

Proof

Recall that

$$\Delta_{\text{gen}}(\hat{\theta}_N) = P\ell_{\hat{\theta}_N} - P_N \ell_{\hat{\theta}_N} = -(P_N - P)\ell_{\hat{\theta}_N}.$$

Taylor expanding $\ell_{\hat{\theta}_N}$ around θ^* gives

$$\ell_{\hat{\theta}_N}(z) = \ell_{\theta^*}(z) + \dot{\ell}_{\theta^*}(z)^\top(\hat{\theta}_N - \theta^*) + o_P(N^{-1/2}).$$

Hence

$$(P_N - P)\ell_{\hat{\theta}_N} = (P_N - P)\ell_{\theta^*} + (P_N - P)\dot{\ell}_{\theta^*}^\top(\hat{\theta}_N - \theta^*) + o_P(N^{-1}).$$

Substituting Eq. 23,

$$\Delta_{\text{gen}}(\hat{\theta}_N) = -(P_N - P)\ell_{\theta^*} + \big[(P_N - P)\dot{\ell}_{\theta^*}\big]^\top H^{-1}\big[(P_N - P)\dot{\ell}_{\theta^*}\big] + o_P(N^{-1}). \tag{24}$$

The first term is $O_P(N^{-1/2})$, while the quadratic term is $O_P(N^{-1})$.Therefore,

$$N\,\mathrm{Var}\big(\Delta_{\text{gen}}(\hat{\theta}_N)\big) \to \mathrm{Var}_P\big(\ell(Z, \theta^*)\big).$$

■

I.2. Proof of Proposition 21

Proof Taking expectations in Eq. 24, the linear empirical-process term vanishes:

$$\mathbb{E}\big[(P_N - P)\ell_{\theta^*}\big] = 0.$$

Thus

$$\mathbb{E}\big[\Delta_{\text{gen}}(\hat{\theta}_N)\big] = \mathbb{E}\left[\big((P_N - P)\dot{\ell}_{\theta^*}\big)^\top H^{-1}\big((P_N - P)\dot{\ell}_{\theta^*}\big)\right] + o(N^{-1}).$$

Now,

$$(P_N - P)\dot{\ell}_{\theta^*} = \frac{1}{N}\sum_{i=1}^{N}\dot{\ell}_{\theta^*}(Z_i),$$

and since

$$P\dot{\ell}_{\theta^*} = 0,$$

cross terms vanish under expectation. Therefore,

$$\mathbb{E}\left[\big((P_N - P)\dot{\ell}_{\theta^*}\big)^\top H^{-1}\big((P_N - P)\dot{\ell}_{\theta^*}\big)\right] = \frac{1}{N}\mathbb{E}\left[\dot{\ell}_{\theta^*}(Z)^\top H^{-1}\dot{\ell}_{\theta^*}(Z)\right].$$

Hence

$$N\mathbb{E}\big[\Delta_{\text{gen}}(\hat{\theta}_N)\big] \to \mathbb{E}\big[K_0(Z, Z)\big].$$

■

Appendix J. For Square-Loss Mean Estimators, the Taylor Series to Second Order in ϵ Is Exact

Let the loss function be of the form $l(z, \theta) := \frac{1}{2}(z - \theta)^T A(z - \theta) := \frac{1}{2}\|z - \theta\|_A^2$, where A is a positive definite matrix. Minimizing $L(\theta, \hat{P}(\epsilon)) + \frac{1}{2}\lambda\,\|\theta - \theta_{\text{ref}}\|^2$, we get, with $\bar{Z}_N := \frac{1}{N}\sum_{i=1}^{N} Z_i$,

$$\begin{aligned}\hat{\theta}(\epsilon) &= (A + \lambda\mathbb{I})^{-1}(A\mathbb{E}_{Z\sim\hat{P}(\epsilon)}[Z] + \lambda\theta_{\text{ref}}) = (A + \lambda\mathbb{I})^{-1}\big[(1 - \epsilon)A\mathbb{E}_{Z\sim P}[Z] + \epsilon A\bar{Z}_N + \lambda\theta_{\text{ref}}\big] \\ &= (1 - \epsilon)\theta^* + \epsilon\hat{\theta}, \end{aligned} \tag{25}$$

namely, $\hat{\theta}(\epsilon)$ interpolates linearly between θ^* and $\hat{\theta} = \hat{\theta}_N(Z_1, \ldots, Z_N)$. Using this fact, we can see that the loss expectations are quadratic in ϵ as the following proposition shows.

Proposition 22

$$\begin{aligned}\mathbb{E}_{\hat{P}}[L(\hat{\theta}(\epsilon), \hat{P}(\epsilon'))] =& \frac{1}{2}\mathbb{E}_{(Z,Z_1,\ldots,Z_N)\sim P^{\otimes(n+1)}}\big[\|Z - (1 - \epsilon)\theta^* - \epsilon\hat{\theta}_N\|_A^2\big] \\ &+ \epsilon'\epsilon\text{Cov}_{(Z_1,\ldots,Z_N)\sim P^{\otimes n}}\Big(A(\bar{Z}_N - \theta^*), \hat{\theta}_N - \theta^*\Big), \end{aligned} \tag{26}$$

where $\hat{\theta}_N$ stands for $\hat{\theta}_N(Z_1, \ldots, Z_N)$.

By setting $\epsilon' = 0, 1, \epsilon$, we can recover the expectation of various average losses of interest. Note that $\mathbb{E}_{\hat{P}}[L(\hat{\theta}(\epsilon), \hat{P}(\epsilon')]$ is a quadratic polynomial in ϵ, ϵ', and therefore, it remains at most quadratic in ϵ, if we set $\epsilon' = 0, 1$, or ϵ. Thus, $\mathbb{E}_{\hat{P}}[\Delta_{\epsilon-\text{Gen}}(\hat{\theta}(\epsilon))]$ and $\mathbb{E}_{\hat{P}}[\Delta R_\lambda(\hat{\theta}(\epsilon))]$ are at most quadratic in ϵ, making a Taylor series in ϵ to second order exact for both quantities.