SI251 - Convex Optimization homework 2

Deadline: 2022-11-16 23:59:59

1. You can use Word, Latex or handwriting to complete this assignment. If you want to submit a handwritten version, scan it clearly.
2. The **report** has to be submitted as a PDF file to Gradescope, other formats are not accepted.
3. You have to write your assignment in English, otherwise you will get a 50% penalty of your score.
4. The submitted file name is **student_id+your_student_name.pdf**.
5. Late policy: You have 4 free late days for the quarter and may use up to 2 late days per assignment with no penalty. Once you have exhausted your free late days, we will deduct a late penalty of 25% per additional late day. Note: The timeout period is recorded in days, even if you delay for 1 minute, it will still be counted as a 1 late day.
6. You are required to follow ShanghaiTech's academic honesty policies. You are not allowed to copy materials from other students or from online or published resources. Violating academic honesty can result in serious sanctions.

Any plagiarism will get Zero point.

1. (20 pts) **Subgradient Methods**. Consider the following problem

$$\begin{array}{ll} \min & f(\mathbf{x}) \\ \text{s.t.} & g_i(\mathbf{x}) \leq 0, \quad i = 1, 2, \ldots, m, \end{array} \tag{1}$$

where $f, g_1, g_2, \ldots, g_m : \mathbb{E} \to \mathbb{R}$ are real-valued convex functions. Please prove the following problems.

(1) Let x^* be an optimal solution of (1), and assume that there exists $\bar{x}$, that satisfies $g_i(\bar{x}) < 0$ for all $g_i(x)$. Then there exist $\lambda_1, \ldots, \lambda_m \geq 0$ for which

$$\mathbf{0} \in \partial f(\mathbf{x}^*) + \sum_{i=1}^{m} \lambda_i \partial g_i(\mathbf{x}^*) \tag{2}$$

$$\lambda_i g_i(\mathbf{x}^*) = 0, \quad i = 1, 2, \ldots, m. \tag{3}$$

(2) If $\mathrm{x}^* \in \mathbb{E}$ satisfies conditions (2) and (3) for some $\lambda_1, \lambda_2, \ldots, \lambda_m \geq 0$, then it is an optimal solution of problem (1).

Solution:

(1) By the Fritz-John conditions there exist $\tilde{\lambda}_0, \tilde{\lambda}_1, \cdots, \tilde{\lambda}_m \geq 0$, not all zeros, for which

$$\mathbf{0} \in \tilde{\lambda}_0 \partial f(\mathbf{x}^*) + \sum_{i=1}^{m} \tilde{\lambda}_i \partial g_i(\mathbf{x}^*) \tag{4}$$

$$\tilde{\lambda}_i g_i(\mathbf{x}^*) = 0, \quad i = 1, 2, \ldots, m \tag{5}$$

We will show that $\tilde{\lambda}_0 \neq 0$. Assume by contradiction that $\tilde{\lambda}_0 = 0$. Then,

$$\mathbf{0} \in \sum_{i=1}^{m} \tilde{\lambda}_i \partial g_i(\mathbf{x}^*) \tag{6}$$

that is, there exist $\boldsymbol{\xi}_i \in \partial g_i(\mathbf{x}^*), i = 1, 2, \ldots, m$, such that

$$\sum_{i=1}^{m} \tilde{\lambda}_i \xi_i = \mathbf{0}$$

Let $\overline{\mathbf{x}}$ be a point satisfying Slater's condition. By the subgradient inequality employed on the pair of points $\overline{\mathbf{x}}, \mathbf{x}^*$ w.r.t. the functions $g_i, i = 1, 2, \ldots, m$, we have

$$g_i(\mathbf{x}^*) + \langle \xi_i, \overline{\mathbf{x}} - \mathbf{x}^* \rangle \leq g_i(\overline{\mathbf{x}}), \quad i = 1, 2, \ldots, m.$$

Multiplying the ith inequality by $\tilde{\lambda}_i \geq 0$ and summing over $i = 1, 2, \ldots, m$ yields

$$\sum_{i=1}^{m} \tilde{\lambda}_i g_i(\mathbf{x}^*) + \left\langle \sum_{i=1}^{m} \tilde{\lambda}_i \boldsymbol{\xi}_i, \overline{\mathbf{x}} - \mathbf{x}^* \right\rangle \leq \sum_{i=1}^{m} \tilde{\lambda}_i g_i(\overline{\mathbf{x}}), \quad i = 1, 2, \ldots, m.$$

Using 5 and 6, we obtain the inequality $\sum_{i=1}^{m} \tilde{\lambda}_i g_i(\overline{\mathbf{x}}) \geq 0$, which is impossible since $\tilde{\lambda}_i \geq 0$ and $g_i(\overline{\mathbf{x}}) < 0$ for any i , and not all the $\tilde{\lambda}_i$'s are zeros. Therefore, $\tilde{\lambda}_0 > 0$, and we can thus divide both the relation 4 and the equalities (3.100) by $\tilde{\lambda}_0$ to obtain that 2 and 3 are satisfied with $\lambda_i = \frac{\tilde{\lambda}_i}{\tilde{\lambda}_0}, i = 1, 2, \ldots, m$.

(2) Suppose then that x^* satisfies 2 and 3 for some nonnegative numbers $\lambda_1, \lambda_2, \ldots, \lambda_m$. Let $\hat{\mathbf{x}} \in \mathbb{E}$ be a feasible point of 1, meaning that $g_i(\hat{\mathbf{x}}) \leq 0, i = 1, 2, \ldots, m$. We will show that $f(\hat{\mathbf{x}}) \geq f(\mathbf{x}^*)$. Define the function

$$h(\mathbf{x}) = f(\mathbf{x}) + \sum_{i=1}^{m} \lambda_i g_i(\mathbf{x}).$$

The function h is convex, and the condition 2 along with the sum rule of subdifferential calculus yields the relation

$$\mathbf{0} \in \partial h\left(\mathrm{x}^*\right),$$

which by Fermat's optimality condition implies that x^* is a minimizer of h over $\mathbb{E}$. Combining this fact with 3 implies that

$$f\left(\mathrm{x}^*\right) = f\left(\mathrm{x}^*\right) + \sum_{i=1}^{m} \lambda_i g_i\left(\mathrm{x}^*\right) = h\left(\mathbf{x}^*\right) \leq h(\hat{\mathbf{x}}) = f(\hat{\mathbf{x}}) + \sum_{i=1}^{m} \lambda_i g_i(\hat{\mathbf{x}}) \leq f(\hat{\mathbf{x}}),$$

where the last inequality follows from the facts that $\lambda_i \geq 0$ and $g_i(\hat{\mathbf{x}}) \leq 0$ for $i = 1, 2, \ldots, m$. We have thus proven that x^* is an optimal solution.

2. (20 pts) **L-smooth functions**. Suppose the function $f : \mathbb{R}^n \to \mathbb{R}$ is convex and differentiable. Please prove that the following relations holds for all $x, y \in \mathbb{R}$ if f with an L-Lipschitz continuous conditions,

$$[1] \Rightarrow [2] \Rightarrow [3] \Rightarrow [4]$$

[1] $\langle \nabla f(x) - \nabla f(y), x - y \rangle \leq L\|x - y\|^2$,

[2] $f(y) \leq f(x) + \nabla f(x)^T (y - x) + \dfrac{L}{2}\|y - x\|^2$,

[3] $f(y) \geq f(x) + \nabla f(x)^T (y - x) + \dfrac{1}{2L}\|\nabla f(y) - \nabla(x)\|^2, \forall x, y,$

[4] $f(\alpha x + (1 - \alpha)y) \leq \alpha f(x) + (1 - \alpha) f(y) - \dfrac{\alpha(1 - \alpha)}{2L}\|\nabla f(x) - \nabla f(y)\|^2, \quad \alpha \in [0, 1].$

Solution:

[1] $\Rightarrow$ [2]: Define the function $G : [0, 1] \to \mathbb{R}$

$$G(t) := f(x + t(y - x)) - f(x) - \langle \nabla f(x), t(y - x) \rangle,$$

so that $G(0) = 0$ and $G(1) = f(y) - f(x) - \langle \nabla f(x), y - x \rangle$. By the fundamental theorem of calculus, we have

$$\begin{aligned} G(1) - G(0) = \int_0^1 G'(t) dt &= \int_0^1 \langle \nabla f(x + t(y - x)) - \nabla f(x), y - x \rangle dt \\ &= \int_0^1 \langle \nabla f(x + t(y - x)) - \nabla f(x), t(y - x) \rangle \frac{1}{t} dt \\ &\leq L\|y - x\|_2^2 \int_0^1 t dt \\ &= \frac{L}{2}\|y - x\|_2^2. \end{aligned}$$

[2] ⇒ [3]: We begin with a useful auxiliary lemma:

Lemma 1. *Consider a differentiable function $g : \mathbb{R}^d \to \mathbb{R}$ satisfying condition [2] and with its global minimum achieved at some v^*. Then*

$$g(v) - g(v^*) \geq \frac{1}{2l}\|\nabla g(v)\|_2^2 \quad \textit{for all } v \in \mathbb{R}^d.$$

Proof: We have

$$\begin{aligned} g(v^*) = \inf_{u \in \mathbb{R}^d} g(u) &\leq \left\{ g(v) + \langle \nabla g(v), u - v \rangle + \frac{L}{2}\|v - u\|_2^2 \right\} \\ &= g(v) - \frac{1}{L}\|\nabla g(v)\|_2^2, \end{aligned}$$

where the last step follows by showing that the minimum of the quadratic program over u is achieved at $u^* = v - \frac{1}{L}\nabla g(v)$, and then performing some algebra.

Note: This lemma and its proof are of independent interest, as they show how gradient descent with step size $1/L$ can be thought of as minimizing a linear approximation along with a quadratic regularization term scaled by $L/2$. Let us now show that [2] ⇒ [3]. For a fixed $x \in \mathbb{R}^d$, define the function

$$g_x(z) = f(z) - \langle \nabla f(x), z \rangle.$$

Note that g_x is convex, differentiable and minimized when $z = x$, and it satisfies our smoothness condition. Hence, the preceding lemma with $v^* = x$ and $v = y$ implies that

$$g_x(y) - g_x(x) \geq \frac{1}{2L}\|\nabla g_x(y)\|_2^2 = \frac{1}{2L}\|\nabla f(y) - \nabla f(x)\|_2^2.$$

A little bit of calculation shows that

$$g_x(y) - g_x(x) = f(y) - f(x) - \langle \nabla f(x), y - x \rangle,$$

which completes the proof.

[3] ⇒ [4]: Let $z = \alpha x + (1 - \alpha)y \in \mathbb{R}^n$, we have

$$\begin{aligned} f(x) &\geq f(z) + \nabla f(z)^T(x - z) + \frac{1}{2L}\|\nabla f(x) - \nabla f(z)\|^2, \\ f(y) &\geq f(z) + \nabla f(z)^T(y - z) + \frac{1}{2L}\|\nabla f(y) - \nabla f(z)\|^2. \end{aligned}$$

Multiplying the first inequality with α and the second inequality with $1 - \alpha$, and adding them together yields

$$\begin{aligned} f(\alpha x + (1-\alpha)y) &\leq \alpha f(x) + (1-\alpha) f(y) - \frac{\alpha}{2L}\|\nabla f(x) - \nabla f(z)\|^2 - \frac{1-\alpha}{2L}\|\nabla f(y) - \nabla f(z)\|^2 \\ &\leq \alpha f(x) + (1-\alpha) f(y) - \frac{\alpha(1-\alpha)}{2L}\|\nabla f(x) - \nabla f(y)\|^2 \end{aligned}$$

where the second inequality follows from the inequality $\alpha\|x\|^2 + (1-\alpha)\|y\|^2 \geq \alpha(1-\alpha)\|x - y\|^2$.

3. **Mirror Descent Methods**.

(1) (15 pts) Let φ be proper convex and differentiable. Suppose $\mathbf{y} = \nabla\varphi(\mathbf{x})$, from conjugate subgradient theorem, show that

$$\varphi(\mathbf{x}) + \varphi^*(\mathbf{y}) = \langle \mathbf{x}, \mathbf{y} \rangle,$$

where $\varphi^*(\mathbf{y})$ is Fenchel conjugate of φ.

Hint: The conjugate subgradient theorem states that if f is closed proper convex, then the following statements are equivalent for a pair of vectors $(\mathbf{x}, \mathbf{y})$: (i) $\langle \mathbf{x}, \mathbf{y} \rangle = f(\mathbf{x}) + f^*(\mathbf{y})$; (ii) $\mathbf{y} \in \partial f(\mathbf{x})$; (iii) $\mathbf{x} \in \partial f^*(\mathbf{y})$, where $f^*(\mathbf{y})$ is Fenchel conjugate of f.

(2) (15 pts) Using Bregman divergence, show that mirror descent has an alternative form, which reads

$$\mathbf{x}^{t+1} = \nabla\varphi^*(\nabla\varphi(\mathbf{x}^t) - \eta_t \mathbf{g}^t),$$

cwhere $\varphi^*(\mathbf{x})$ is Fenchel conjugate of φ, $\mathbf{g}^t \in \partial f(\mathbf{x}^t)$ and $\eta_t > 0$ is the stepsize. (For simplicity, assume the constraints set $\mathcal{C} = \mathbb{R}^n$.)

Solution:

(1) Since φ is convex and $\mathbf{y} = \nabla\varphi(\mathbf{x})$, we have

$$\begin{gathered}
\varphi(\mathbf{m}) \geq \varphi(\mathbf{x}) + \langle \mathbf{y}, \mathbf{m} - \mathbf{x} \rangle \\
\Updownarrow \\
\langle \mathbf{y}, \mathbf{x} \rangle - \varphi(\mathbf{x}) \geq \langle \mathbf{y}, \mathbf{m} \rangle - \varphi(\mathbf{m}) \\
\Updownarrow \\
\langle \mathbf{y}, \mathbf{x} \rangle - \varphi(\mathbf{x}) \geq \sup_{\mathbf{m}} \langle \mathbf{y}, \mathbf{m} \rangle - \varphi(\mathbf{m}) \\
\Updownarrow \\
\langle \mathbf{y}, \mathbf{x} \rangle - \varphi(\mathbf{x}) \geq \varphi^*(\mathbf{y})
\end{gathered}$$

On the other hand

$$\begin{aligned}
\varphi^*(\mathbf{y}) &= \sup_{\mathbf{x}} \langle \mathbf{y}, \mathbf{x} \rangle - \varphi(\mathbf{x}) \\
&\geq \langle \mathbf{y}, \mathbf{x} \rangle - \varphi(\mathbf{x})
\end{aligned}$$

Therefore, arranging the term, we have $\varphi(\mathbf{x}) + \varphi^*(\mathbf{y}) = \langle \mathbf{x}, \mathbf{y} \rangle$. This completes the proof.

(2)

$$\begin{aligned}
\mathbf{x}^{t+1} &= \arg\min_{\mathbf{x} \in \mathcal{C}} \left\{ \left\langle \mathbf{g}^t, \mathbf{x} - \mathbf{x}^t \right\rangle + \frac{1}{\eta_t} \left(\varphi(\mathbf{x}) - \varphi\left(\mathbf{x}^t\right) - \left\langle \nabla\varphi\left(\mathbf{x}^t\right), \mathbf{x} - \mathbf{x}^t \right\rangle \right) \right\} \\
&= \arg\min_{\mathbf{x} \in \mathcal{C}} \left\{ \left\langle \mathbf{g}^t - \frac{1}{\eta_t} \nabla\varphi\left(\mathbf{x}^t\right), \mathbf{x} - \mathbf{x}^t \right\rangle + \frac{1}{\eta_t} \left(\varphi(\mathbf{x}) - \varphi\left(\mathbf{x}^t\right) \right) \right\} \\
&= \arg\min_{\mathbf{x} \in \mathcal{C}} \left\{ \left\langle \mathbf{g}^t - \frac{1}{\eta_t} \nabla\varphi\left(\mathbf{x}^t\right), \mathbf{x} - \mathbf{x}^t \right\rangle + \frac{1}{\eta_t} \varphi(\mathbf{x}) \right\} \\
&= \arg\min_{\mathbf{x} \in \mathcal{C}} \left\{ \left\langle \eta_t \mathbf{g}^t - \nabla\varphi\left(\mathbf{x}^t\right), \mathbf{x} - \mathbf{x}^t \right\rangle + \varphi(\mathbf{x}) \right\} \\
&= \arg\max_{\mathbf{x} \in \mathcal{C}} \left\{ \left\langle \nabla\varphi\left(\mathbf{x}^t\right) - \eta_t \mathbf{g}^t, \mathbf{x} - \mathbf{x}^t \right\rangle - \varphi(\mathbf{x}) \right\}
\end{aligned}$$

Therefore, we have $x^{t+1} = \nabla\varphi^*\left(\nabla\varphi\left(x^t\right) - \eta_t g^t\right)$

4. **Natural gradient descent**. Natural gradient descent is an optimization method motivated by the difference between the parameter space (e.g., parameters of a neural network) and the distribution space of the output (e.g., categorical distribution given by a classifier), and performs well for many applications as an alternative to stochastic gradient descent. Now consider a simple classification problem as below.

$$\min_{\theta} \mathbb{E}_{x,y_{true}\sim\rho_{data}} \left[\mathcal{L}(p(\cdot|x;\theta), y_{true})\right] \tag{7}$$

where x and y_{true} are the feature and label of the sample respectively, ρ_{data} is the distribution of the samples, $\mathcal{L}$ is some loss function, e.g., cross-entropy loss, and $p(\cdot|x;\theta)$ is the classifier that we want to obtain, which outputs the probabilities that the sample x belongs to class ys.

For some reason, we cannot optimize the classification problem above directly. In such case, we will often find a surrogate function to approximate the original problem in each iteration.

However, The surrogate function is just an approximation to the original problem, and there will be a large bias when the updated classifier is far from itself before. Hence, we need to restrict the update within a trust region.

That is what natural gradient descent does. Mathematically, the formulation of it is

$$\begin{aligned} \min_{\theta} \quad & \mathbb{E}_{x,y_{true}\sim\rho_{data}} \left[\mathcal{S}_{\theta_{old}}(p(\cdot|x;\theta), y_{true})\right] \\ \text{subject to} \quad & \mathbb{E}_{x\sim\rho_{data}} \left[\mathrm{kl}\left(p(\cdot|x;\theta_{old}), p(\cdot|x;\theta)\right)\right] \leq \delta \end{aligned} \tag{8}$$

where θ_{old} is the parameters of the classifier $f(x;\theta)$ before the update, $\mathcal{S}_{\theta_{old}}$ is the surrogate function of $\mathcal{L}$ at the point θ_{old}, and the averaged KL divergence

$$\mathbb{E}_{x\sim\rho_{data}} \left[\mathrm{kl}\left(p(\cdot|x;\theta_{old}), p(\cdot|x;\theta)\right)\right] \leq \delta \tag{9}$$

is to restrict the update to the trust region of θ_{old}.

(Hint: $\mathrm{kl}(p,q) = \sum_i p_i \log \frac{p_i}{q_i}$)

The story ends.

(1) (10 pts) Please prove that, when $\theta = \theta_{old}$, $\nabla_\theta \mathbb{E}_{x\sim\rho_{data}} \left[\mathrm{kl}\left(p(\cdot|x;\theta_{old}), p(\cdot|x;\theta)\right)\right] = 0$, which is known as score function in probability and statistics.

(2) (20 pts) To perform natural gradient descent, we should do an approximation as below on the objective and constraint further.

$$\begin{aligned} \min_{\theta} \quad & g^T\theta \\ \text{subject to} \quad & \frac{1}{2}(\theta-\theta_{old})^T H(\theta-\theta_{old}) \leq \delta \end{aligned} \tag{10}$$

where $g = \nabla_\theta \mathbb{E}_{x,y_{true}\sim\rho_{data}} \left[\mathcal{S}_{\theta_{old}}(p(\cdot|x;\theta), y_{true})\right], H = \nabla^2_\theta \mathbb{E}_{x\sim\rho_{data}} \left[\mathrm{kl}\left(p(\cdot|x;\theta_{old}), p(\cdot|x;\theta)\right)\right]$. As we have proved above, there is no need to consider the derivative of the constraint.

Now that please prove that the update formula of natural gradient descent is

$$\theta = \theta_{old} - \sqrt{\frac{2\delta}{g^T H^{-1} g}} H^{-1} g \tag{11}$$

(3) (Bonus 20 pts) In practice, $p(\cdot|x;\theta)$ is usually a complicated model such as neural network. Hence the hessian matrix H of the constraint is hard to compute.

However, the hessian matrix H here is also known as Fisher Information Matrix in the probability and statics, which holds an amazing equality that we can infer H directly just computing the gradient of the log likelihood.

Below is that equality. Please prove it.

$$\begin{aligned}&\nabla_\theta^2 \mathbb{E}_{x\sim\rho_{data}}\left[\mathrm{kl}\left(p(\cdot|x;\theta_{old}), p(\cdot|x;\theta)\right)\right]\\ =\ &\mathbb{E}_{x\sim\rho_{data}}\left[\mathbb{E}_{y\sim p(\cdot|x;\theta_{old})}\left[\nabla_\theta \log\left(p(y|x;\theta)\right)\nabla_\theta \log\left(p(y|x;\theta)\right)^T\right]\right]\end{aligned} \tag{12}$$

Solution:

(1)

$$\begin{aligned}\nabla_\theta \mathbb{E}_{x\sim\rho_{data}}\left[\mathrm{kl}\left(p(\cdot|x;\theta_{old}), p(\cdot|x;\theta)\right)\right] &= \nabla_\theta \mathbb{E}_{x\sim\rho_{data}}\left[-\sum_y p(y|x;\theta_{old})\log(p(y|x;\theta))\right]\\ &= \mathbb{E}_{x\sim\rho_{data}}\left[-\sum_y p(y|x;\theta_{old})\nabla_\theta \log(p(y|x;\theta))\right]\\ &= \mathbb{E}_{x\sim\rho_{data}}\left[-\sum_y p(y|x;\theta_{old})\frac{\nabla_\theta p(y|x;\theta)}{p(y|x;\theta)}\right]\end{aligned}$$

Since $\theta = \theta_{old}$, we have

$$\begin{aligned}\nabla_\theta \mathbb{E}_{x\sim\rho_{data}}\left[\mathrm{kl}\left(p(\cdot|x;\theta_{old}), p(\cdot|x;\theta)\right)\right] &= \mathbb{E}_{x\sim\rho_{data}}\left[-\sum_y \nabla_\theta p(y|x;\theta)\right]\\ &= \mathbb{E}_{x\sim\rho_{data}}\left[-\nabla_\theta \sum_y p(y|x;\theta)\right]\\ &= \mathbb{E}_{x\sim\rho_{data}}\left[-\nabla_\theta 1\right]\\ &= 0\end{aligned}$$

(2) According to KKT optimality condition, we have

$$\begin{aligned}\mathcal{L}(\theta,\lambda) &= g^T\theta + \frac{1}{2}\lambda\left[(\theta-\theta_{old})^T H(\theta-\theta_{old}) - \delta\right]\\ \nabla_\theta \mathcal{L}(\theta,\lambda) &= 0\end{aligned}$$

Then, we have

$$\theta = \theta_{old} - \frac{1}{\lambda}H^{-1}g$$

Applying it to the inequality constraint, we have

$$\frac{1}{2\lambda^2}gH^{-1}g \le \delta$$

Considering complementary slackness, we have $\frac{1}{\lambda} = \sqrt{\frac{2\delta}{gH^{-1}g}}$.

Finally, we have

$$\theta = \theta_{old} - \sqrt{\frac{2\delta}{g^T H^{-1} g}} H^{-1} g$$

(3)

$$\begin{aligned}
&\nabla_\theta^2 \mathbb{E}_{x\sim\rho_{data}} \left[\mathrm{kl}\left(p(\cdot|x;\theta_{old}), p(\cdot|x;\theta)\right)\right] \\
=&\nabla_\theta^2 \mathbb{E}_{x\sim\rho_{data}} \left[-\sum_y p(y|x;\theta_{old}) \log(p(y|x;\theta))\right] \\
=&\mathbb{E}_{x\sim\rho_{data}} \left[\mathbb{E}_{y\sim p(\cdot|x;\theta_{old})} \left[-\nabla_\theta^2 \log(p(y|x;\theta))\right]\right]
\end{aligned}$$

Now, we just consider the term inside the bracket.

$$-\nabla_\theta^2 \log(p(y|x;\theta)) = -\frac{\nabla_\theta^2 p(y|x;\theta)}{p(y|x;\theta)} + \nabla_\theta \log\left(p(y|x;\theta)\right) \nabla_\theta \log\left(p(y|x;\theta)\right)^T$$

Thus, we have

$$\begin{aligned}
&\mathbb{E}_{x\sim\rho_{data}} \left[\mathbb{E}_{y\sim p(\cdot|x;\theta_{old})} \left[-\nabla_\theta^2 \log(p(y|x;\theta))\right]\right] \\
=&\mathbb{E}_{x\sim\rho_{data}} \left[\mathbb{E}_{y\sim p(\cdot|x;\theta_{old})} \left[-\frac{\nabla_\theta^2 p(y|x;\theta)}{p(y|x;\theta)} + \nabla_\theta \log\left(p(y|x;\theta)\right) \nabla_\theta \log\left(p(y|x;\theta)\right)^T\right]\right] \\
=&\mathbb{E}_{x\sim\rho_{data}} \left[\nabla_\theta 1 + \mathbb{E}_{y\sim p(\cdot|x;\theta_{old})} \left[\nabla_\theta \log\left(p(y|x;\theta)\right) \nabla_\theta \log\left(p(y|x;\theta)\right)^T\right]\right] \quad (p(y|x;\theta) = p(y|x;\theta_{old})) \\
=&\mathbb{E}_{x\sim\rho_{data}} \left[\mathbb{E}_{y\sim p(\cdot|x;\theta_{old})} \left[\nabla_\theta \log\left(p(y|x;\theta)\right) \nabla_\theta \log\left(p(y|x;\theta)\right)^T\right]\right]
\end{aligned}$$